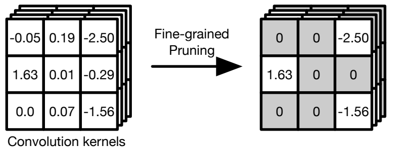
Focused Quantization for Sparse CNNs

Yiren Zhao*[1], Xitong Gao*[2], Daniel Bates[1], Robert Mullins[1], Cheng-Zhong Xu[3]
* Equal contribution [1] University of Cambridge [2] Shenzhen Institutes of Advanced Technology [3] University of Macau

Paper GitHub

Fine-grained Pruning

…provides the **best compression** by removing connections at the finest granularity, *i.e.* individual weights:



Shift Quantization

…or powers-of-two quantization constraints weight values in a model to powers-of-two or zero:

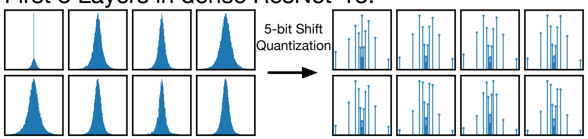
$$\{0, \pm 1, \pm 2, \pm 4, \pm 8, \ldots\}$$

Efficient hardware: multiplications ⟶ bit-shift

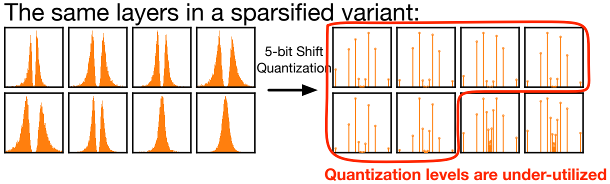
Challenges

Shift quantization is however often **in conflict with** fine-grained pruning:

First 8 Layers in dense ResNet-18:



The same layers in a sparsified variant:

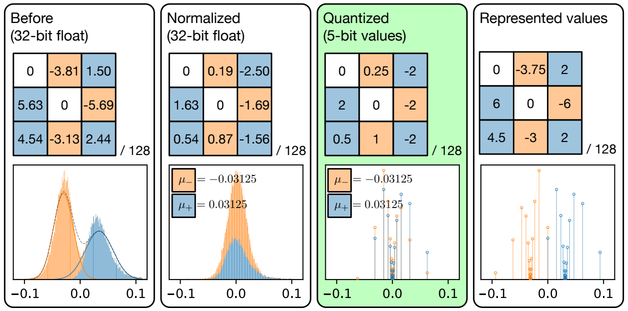


How can we quantize sparse weights ***efficiently*** and ***effectively***?

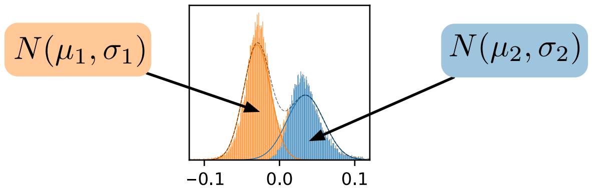
Efficiency: reduced model size, minimized compute cost
Effectiveness: quantization levels are well-utilized: better accuracy.

Focused Quantization

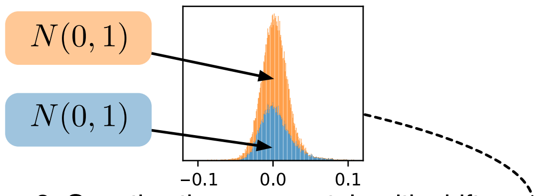
The quantization process for the unpruned weights on `block3f/conv1` in sparse ResNet-50:



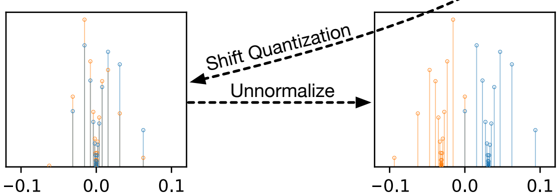
Step 1: Use maximum-likelihood estimation to find out the approximate Gaussian mixture, assign weights to the Gaussian components by sampling:



Step 2: Normalizes the two Gaussian components:

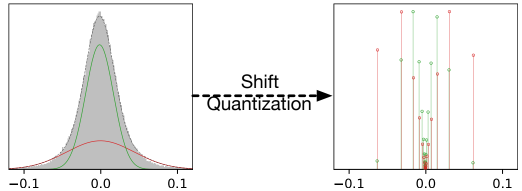


Step 3: Quantize them separately with shift quantization:



Wasserstein Separation

If the two components are close together, i.e. $W(c_1, c_2) = \frac{1}{\sigma^2}\left((\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2\right) \leq w_{\text{sep}}$ we instead use shift quantization.

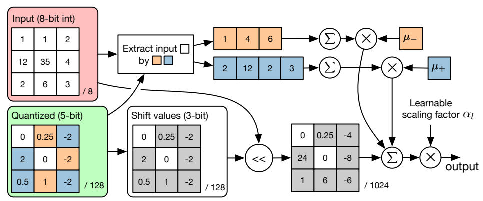


Comparisons to SOTA

ResNet-18	Top-1	Top-5	Size (MB)	CR (×)
TTQ [24]	66.00	87.10	2.92*	16.00*
INQ (2 bits) [23]	66.60	87.20	2.92*	16.00*
INQ (3 bits) [23]	68.08	88.36	4.38*	10.67*
ADMM (2 bits) [14]	67.0	87.5	2.92*	16.00*
ADMM (3 bits) [14]	68.0	88.3	4.38*	10.67*
ABC-Net (5 bases, or 5 bits) [15]	67.30	87.90	7.30*	6.4 *
LQ-Net (preact, 2 bits) [22]	68.00	88.00	2.92*	16.00*
D&Q (large) [19]	**73.10**	**91.17**	21.98	2.13*
Coreset [3]	68.00	—	3.11*	15.00
Focused compression (5 bits, sparse)	68.36	88.45	**2.86**	**16.33**

ResNet-50	Top-1	Top-5	Size (MB)	CR (×)
INQ (5 bits) [23]	74.81	92.45	14.64*	6.40*
ADMM (3 bits) [14]	74.0	91.6	8.78*	10.67*
ThiNet [17]	72.04	90.67	16.94	5.53*
Clip-Q [21]	73.70	—	6.70	14.00*
Coreset [3]	74.00	—	5.93*	15.80
Focused compression (5 bits, sparse)	**74.86**	**92.59**	**5.19**	**18.08**

Efficient Hardware Design



Configuration	#Gates	Ratio
ABC-Net (5 bases, or 5 bits)	806.1 M	2.93×
LQ-Net (2 bits)	314.4 M	1.14×
Shift quantization (3 bits, unsigned)	275.2 M	1.00×
FQ (5 bits)	275.6 M	1.00×
FQ (5 bits) + Huffman	276.4 M	1.00×